# Evaluating Artificial Social Intelligence in an Urban Search and Rescue Task Environment

**Jared Freeman[1], Lixiao Huang[2], Matt Wood[1], Stephen J. Cauffman[2]**

Aptima Inc.[1], Arizona State University[2]

freeman@aptima.com[1], Lixiao.Huang@asu.edu[2], mwood@aptima.com[1], scauffma@asu.edu[2]

**Abstract**

Human team members show a remarkable ability to infer the state of their partners and anticipate their needs and actions. Prior research demonstrates that an artificial system can make some predictions accurately concerning artificial agents. This study investigated whether an artificial system could generate a robust Theory of Mind of human teammates. An urban search and rescue (USAR) task environment was developed to elicit human teamwork and evaluate inference and prediction about team members by software agents and humans. The task varied team members' roles and skills, types of task synchronization and interdependence, task risk and reward, completeness of mission planning, and information asymmetry. The task was implemented in Minecraft™ and applied in a study of 64 teams, each with three remotely distributed members. An evaluation of six Artificial Social Intelligences (ASI) and several human observers addressed the accuracy with which each predicted team performance, inferred experimentally manipulated knowledge of team members, and predicted member actions. All agents performed above chance; humans slightly outperformed ASI agents on some tasks and significantly outperformed ASI agents on others; no one ASI agent reliably outperformed the others; and the accuracy of ASI agents and human observers improved rapidly though modestly during the brief trials.

## Introduction

Teams succeed through coordinated actions by members who differ in their capabilities and roles. Teams achieve this coordination using domain-specific, often mission-specific, compositions of training, talk, technology, and Theory of Mind (ToM; Baron-Cohen, Leslie, and Frith 1985; Premack and Woodruff 1978). The first three techniques help team members manage the scope of the fourth, ToM, which enables team members to infer the capabilities and goals of teammates, predict their actions, and coordinate or compensate to improve teamwork.

Training—by which we mean education, planning, rehearsal, and repeated mission execution—demonstrably develops members' ability to predict the responses of colleagues to challenges, coordinate actions, and improve performance (McNeese et al. 2015; Yurko et al. 2020). Such preparation is essential over all missions but typically inadequate for any given mission because training cannot perfectly anticipate a specific mission (nor should it if it is to ensure generalizability of learning across potential variants of the mission).

To compensate for the inadequacy of training, teams communicate in real time using spoken language (e.g., military air control communications protocols) and symbolic language (e.g., the marking conventions used by search and rescue teams). These communication protocols range from the formal to the informal. They convey among team members much of what any member knows, needs, or intends at the moment. However, communications are often inaccurate, incomplete, untimely, or unavailable (e.g., in military operations).

Technologies designed to improve team coordination can compensate for the inevitable inadequacies of training and talk. MacMillan et al. (2002) describe applications of multi-objective optimization, simulation, and empirical research to design teams that are better sized and synchronized, and thus outperform standard teams in empirical and computational studies. The rich literatures of operations research and robotics describe techniques for making plans more efficient, robust, or resilient, and planning more rapidly (c.f., Kitano et al., 1997). Real-time social network analysis and related techniques have been used to help military commanders assess and improve

teamwork in their organizations in near real time (Brown et al. 2017). In many military operations, sophisticated technologies represent the state of a mission on displays and recommend or enact responses to threats, thus automating what would otherwise require human coordination (c.f., Aegis doctrine for automatically executing tactical actions defined by policy). Such technologies are, however, not available in all domains at all times, nor competent or trusted by their users in all situations.

Through training, talk, and technology, team members scope, develop, and maintain mental models of their teammates, or Theory of Mind (ToM). They use these models when training, talk, and technology are insufficient to coordinate with team members. More specifically, ToM enables team members to infer the cognitive and affective state of teammates, their goals, and their needs; to predict teammates' actions; and to develop guidance and actions that coordinate work. Such inference is ineluctable, difficult, and errorful.

Recent research explores whether we can develop a Machine Theory of Mind (MToM) to offload from humans some of the burden of developing, validating, maintaining, and applying ToM of teammates. Rabinowitz et al. (2018) demonstrated that meta-learning could be used by an artificial system to make accurate inferences and predictions concerning artificial agents. Such work, if extended to model multiple human team members, might be called Artificial Social Intelligence (ASI) or Machine Theory of Teams (MToT). ASI could potentially generate advice that improves teams in the most difficult of circumstances, those in which team members are highly varied in their capabilities and capacity, task synchronization is complex, there is risk of failure at high stakes, preparation and information are incomplete, communication is encumbered, and thus the necessity and difficulty of building ToM is high.

Research to develop ASI requires collaboration between computer scientists and social scientists. DARPA sponsors this research in a program called Artificial Social Intelligence for Successful Teams (ASIST). ASIST engages six teams in the development of ASI; six teams in social science research intended to improve ASI inference, prediction, and intervention; and one team (our own) that is focused on experimental design, testbed development, and evaluation. Below, we briefly describe the design of the most recent study to advance ASIST (see Huang et al. 2021 for details) and the initial findings from evaluation of the accuracy[2] of MToM applied to human teams.

# Method

## Experiment Design and Task Environment

Development and evaluation of ASI require a team task that systematically demands and complicates the generation of ToM. The domain of Urban Search and Rescue (USAR) instantiates many of the attributes of such a team task (enumerated above and in Table 1).

| Team Attri-butes | USAR Features | Minecraft ASIST Task Features |
|---|---|---|
| Skills & roles | USAR requires heterogeneous teams (different roles); individual and team tasks | Three team members and three possible roles; individual and team tasks |
| Task synchronizati on | Temporal constraints with asynchronous, sequential, and simultaneous tasks | 15 min missions to rescue victims in tasks requiring asynchronous, sequential, and simultaneous teamwork |
| Risk | Hazards, unexpected events for workers | Hidden freeze plate in rooms |
| Reward | Some victims more severe, triage necessary | Critical victims vs. regular victims (high vs. low reward) |
| Preparation | Planning is important | Planning session was manipulated |
| Information | Incomplete information on victim and blockage location | Maps are incomplete |
| Communicat ion | Verbal comms & searched areas marked to communicate with team | Audio & marker blocks for comms (Divergent marker keys create conflicting mental models) |

Table 1: Comparison of teamwork features in USAR and a Minecraft USAR simulation.

[2] We are agnostic to the genesis of ToM in humans, whether it be an innate capability to apply theory (the "theory theory" account of Gopnik, 1992) or mental simulation (Gordon, 1986) to infer and predict. Training, talking, and technology, discussed above, are data sources for both mechanisms, whether building theory from empirical data, or specifying input conditions and operating rules for simulation. The genesis of human ToM may be an inspiration for the design of MToM, but it is not a constraint assessment of accuracy, which is our current focus.

A USAR task environment was built using Minecraft, which provides a lightweight method to simulate the task constraints that a typical USAR team might encounter (Corral et al. in press; Lematta et al. 2019). A study was designed in which teams of three were tasked with finding and rescuing victims of two types, regular and critical, in 15-min missions. Each team member could choose from three roles (a searcher—finding and relocating victims, a medic—treat victims, and an engineer—removing rubble). Task costs and rewards varied: regular victims could be rescued by a single medic in 7.5 seconds for 10 points each; critical victims required three team members to join together to "wake" the victim from a coma, after which a medic could rescue the victim in 15 seconds for 50 points each. A few threat rooms had hidden freeze plates that could immobilize any team member who enters the room and require rescue by a medic teammate. Each of 64 teams of remotely distributed members executed two different missions in the same virtual building but with different victim and rubble layouts. Mission order was counterbalanced between teams.

To control the degree to which the team shared a model of roles, goals, and strategy at the start of missions, we manipulated the opportunity to plan together. 32 teams in a planning condition were given three minutes to develop a plan for approaching the task before the second trial. 32 teams in a control condition performed math problems for three minutes instead.

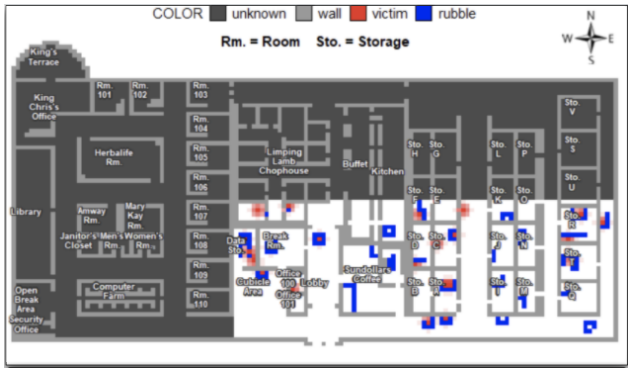


Figure 1: An example map depicts building structure and, in two white areas, victims and rubble. Maps for two other players reveal victims and rubble in the left bottom and central bottom, and in the central top and bottom top.

To control the distribution of starting knowledge over team members, we provided maps that displayed one building section to all members, a unique section to each member, and hid sections from all members (see Figure 1). The maps accurately showed locations of fallen rubble and victims.

To control the agreement of knowledge between team members, we provided conflicting definitions (or legends) for three marker blocks carried by all members to lay as signals to others; the colors of marker blocks designate the ownership (e.g., participant "Red" lays red marker blocks), but the symbols (1, 2, and 3) on marker blocks are the same between team members. Two members received a legend in which the marker block labelled 1 meant "no victim here", marker block 2 meant "regular victim here". One member received the same marker blocks, but the definitions were reversed in the legend (see Figure 2). All members received a marker block 3 defined as "critical victim here." This conflict in marker block meaning enabled us to apply a variant of the "Sally-Anne" test of Theory of Mind, in which the ASI must infer whether one member will enter a room given potentially false beliefs about the meaning of another's marker block (Rabinowitz et al. 2018; Wimmer and Perner 1983). Maps and marker block legends were displayed during each trial, to help participants form and execute search and rescue strategies (see Figure 3).

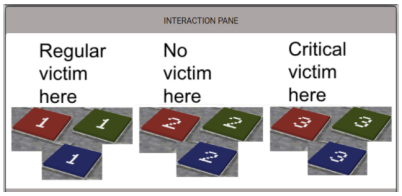


Figure 2: A marker block legend with "reversed" meaning.

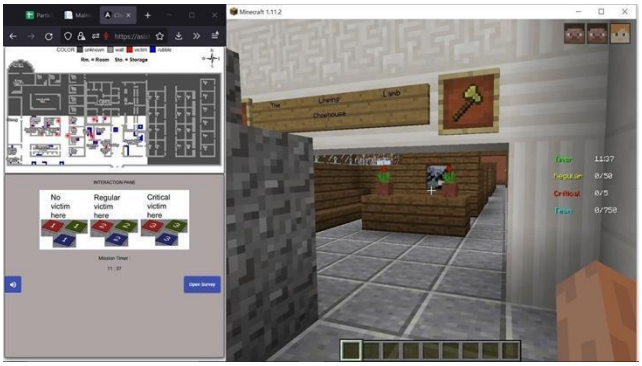

Figure 3: Participant interface displays the information map (top left), marker block legend (bottom left), and Minecraft (right).

In sum, we created a between-within group mixed experimental design. The between-group factor was planning (with or without) before the second trial. The within-group factor was two missions of equivalent difficulty in terms of victim and rubble layout. At the individual level, we also manipulated the three versions of information maps for different known regions, and two

versions of marker block legends. Individual tasks and interdependent tasks were designed to allow us to study and ASI to model individual taskwork, two-member coordination (in response to threat rooms), three-member coordination (to rescue critical victims), as well as individual and team navigation and rescue strategies.

## Participants

201 participants (67 teams) were recruited from Reddit, Discord, and University listserv with the requirement of playing Minecraft, living in the United States, speaking English, and having a normal color vision. Three teams' data were omitted due to flaws in displaying the information map and marker blocks). The remaining 64 teams (192 participants) consisted of 141 males, 49 females, and 2 individuals who declared other gender identities or preferred not to respond. The mean age of participants was 22.04 (*SD*=5.22, ranging from 18 to 49). The most common ethnicities were white/Caucasian (54.2%; 104), Asian (25.8%; 49), and Hispanic or Latino (13%; 25). All participants had at least a high school level education.

## Procedure

The remotely conducted two-session experiment lasted for 3.5 hours. In Session 1 (one hour), participants checked in to install the required software (e.g., Minecraft, Forge mods, and Zoom™) correctly and then finished pre-dispositional surveys. In Session 2 (which ran two and a half hours), three qualified participants were required to join as a team to go through a voice-over training video, a hands-on practice of required individual actions and team interactions in the Minecraft™ world, an independent action-based Minecraft competency test, and then two formal missions. Participants filled out survey sections after each step in Session 2. Consent forms were attained at the beginning.

## Data, Metrics, and Measures

The study used 469 survey items to elicit or quantify 22 constructs spanning demographics, individual differences (e.g., personality, spatial ability, game experience), accuracy of Theory of Mind, and teamwork process and climate. The testbed message bus registered all experimental metadata (e.g., the specific assignment of marker block legends and maps to participants, and identifiers of trials and teams) and all events in the Minecraft world (e.g., moving, using tools, and rescuing a victim), their timestamps, locations, and the entities involved. In addition, the study captured experimenters' Bird's-eye view videos of trials (see Figure 4); Zoom™ videos, audio, and transcriptions; as well as various abstractions of the data (e.g., field of view, semantic translation of location, speech act classes).

ASI agents generated measurements relevant to four metrics of accuracy (see Table 2) using various combinations of the data above. Human observers generated measurements on the four metrics from Bird's-eye view videos that presented participants' locations on the building map, their first-person views, and their voice communication. Both ASI agents and human observers' performance was evaluated over 24 trials (18.75% of 128 experimental trials) that were held out for use in testing ASI (thus not available for training ASI).


Figure 4: Bird's-eye view displays Zoom speakers' identity and participants' game views (top) and the mission space (bottom).

Both humans and ASI agents were evaluated on the same four metrics (see Table 2) relevant to MToM. The accuracy of ASI prediction of effects (specifically, ASI predictions of the team's terminal game score) was measured at three time points per trial (i.e., min 4, 9, and 14), on the theory that ASI must estimate the effects of its potential interventions to select well between them. The accuracy of ASI inferences concerning participant knowledge was measured because such inferences are presumably required by ASI to generate interventions that remedy differences in accurate knowledge between team members (here, M3 concerned systematic differences in the content of accurate but incomplete maps provided to participants) and conflicting or false beliefs (M6 addressed experimentally induced conflicts in beliefs between participants concerning the meaning of numbered marker blocks). Measurements on M3 and M6 were also taken at three time points per trial. The accuracy of ASI predictions of actions based on conflicting beliefs between participants was measured in a variant of the Sally-Anne task in which ASI predicted whether a participant would enter a room given the presence of a marker block laid by a participant with a conflicting belief about that block's meaning. A measurement was computed at each such event for the one player whose block semantics differed from that of the

other two players. We note that the number of evaluation metrics (four) was intentionally small. This ensured that ASI developers invested a modest effort in satisfying program assessment requirements but had spare capacity to create and test alternative MToM capabilities.

| Metric ID: Function | ASI Agent & Human Observer infer / predict | Measure |
|---|---|---|
| M1: Prediction of effects of future interventions | Team score (3x per trial at fixed times) | Normalized RMSE, or RMSE divided by the mean of all team scores for each trial. 1–NRMSE presented here to match the direction of other measures. |
| M3: Inference of member mental model / knowledge | Given map information (3x) | Mean accuracy. Guessing computed as the conditional probability of correctly guessing the first then second of three map options in each trial, or ⅓ x ½ = ⅙ |
| M6: Inference of member mental model / knowledge | Given marker block meanings (3x) | Mean accuracy. Guessing computed as conditional probability of correctly guessing the first then second options. Value is ⅓ since one of two labels is twice as frequent. |
| M7: Prediction of action given member beliefs (Sally-Anne) | Room entry in response to another participant's marker block (many per trial) | Mean accuracy. Guessing computed based on random selection of 'did enter' or 'did not enter' room for each occurrence. |

Table 2: Metrics for ASI Agents and human Observers.

## Artificial Social Intelligence

Six program performers employed tasked with designing and building ASI created distinctly different agents to process the data from this experiment into inferences and predictions.

The University of Arizona team, led by Adarsh Pyarelal, used dynamic Bayes networks (DBNs) to model individual and team activity states and mental states (ToM), using in-game participant behavior, natural language processing, and speech analysis.

The SIFT team, led by Chris Geib, employed MC Tree Search over learnable action grammars to generate multiple candidate explanations for observed behavior. Explanations included explicit ascriptions of ToM beliefs for each agent. The system then used weighted model counting over the explanations to probabilistically infer the most likely mental states and asymmetric beliefs between team members.

The team from University of Southern California, led by David Pynadath, applied recursive POMDPs as candidate participant models with ToM, constructed by combining a RDDL specification of the domain with perturbations along domain-independent dimensions. The ASI agent performed Bayesian inference to update beliefs over these candidate models based on observed team and individual behavior.

The team from DOLL/MIT, led by Paul Robertson, generated narratives from stories that represent, for each team member, a story of the team. The Narrative provided a rationale for the past and predictions for the future. This ASI agent also used mechanisms for inverse planning, probabilistic ToM, probabilistic conditional preference, story understanding (Genesis), and learned player capability, such as speed.

The team from Carnegie Mellon, led by Katia Sycara, implemented a modular neural network Theory of Mind (ToM) model that infers an individual's beliefs, goals and intentions from observations and environmental context; introspection resolves deviations between predicted and observed behaviors. Combined ToM models of teammates provided reasoning over shared mental models, team processes and produce appropriate individual and team interventions.

The team from Charles River Analytics, led by Bryan Loyall, created a Cognitive Inverter that uses probabilistic programming to recognize goals, behaviors, and mental states from open world observations. A Strategic Coach will select the most effective interventions, based on principles from interactive narrative research.

## Findings

An evaluation compared inferences and predictions by six ASI agents to a human baseline of three observers for M1, M3, and M6, as well as that of two observers for M7, and to a guessing baseline that assumes a random draw from a known distribution for the possible response options. As summarized in Figure 5, the human baseline was higher

(better) than the performance of all ASI agents on each metric, and when computed both as an average rank (human baseline rank = 1.0, average ASI agent rank = 4.2) and in terms of average performance over median value on each measure (human baseline = 0.12, ASI agents = -0.01). No one ASI agent consistently outperformed the others. The variation between agents is likely due to differences in approach. Variation between humans and ASI agents may be due both to differences in their respective inference and prediction methods, and variations in the data that fed those methods. ASI agents consumed testbed message bus data, humans used mainly video and audio. These data sources differ in representation of information and in the information they represent. Humans and ASI agents performed better than guessing in nearly all cases; average performance over guessing was similar between the human baseline (0.47) and ASI agents (0.35) but varied somewhat by metric (see Figure 6).

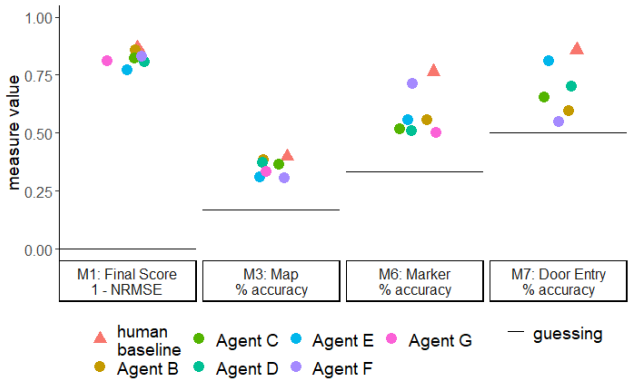


Figure 5: Accuracy of human observers (triangle) and artificial agents (circles) on four tests of social intelligence.

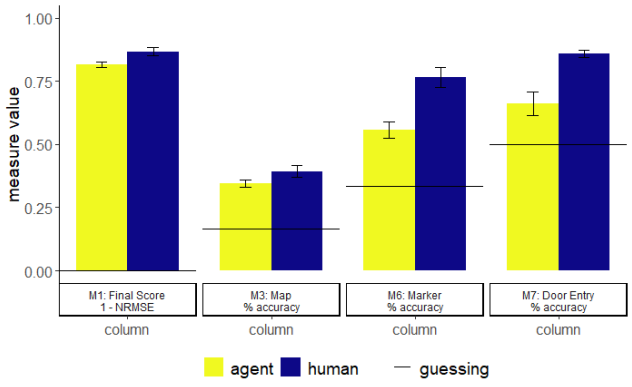


Figure 6: Accuracy of human observers (yellow) and average of artificial agents (blue) on four tests of social intelligence. Error bars where provided represent +/- 1SE.

For those measures on which agents performed most similarly to each other (M1 and M6), agent accuracy tended to improve over time within each trial (see Figures 7 and 8).

The results collectively suggest that these ASI agents were able to reliably predict team score (M1) and actions of individual members (M7), infer divergent beliefs (M3), and infer false beliefs (M6). However, the ability of these ASI agents to infer false beliefs (M6) and predict future actions related to false beliefs (M7) lags further behind human capabilities than their ability to predict future performance (M1) and infer divergent beliefs (M3).

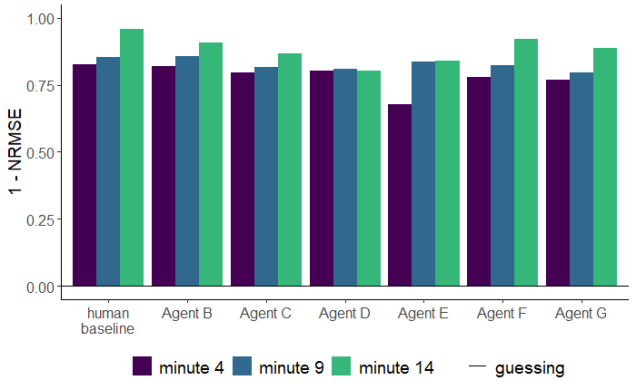


Figure 7: Accuracy predicting final score (M1) thrice per trial, measured as 1-NRMSE. Guessing would result in a score of zero on this measure.

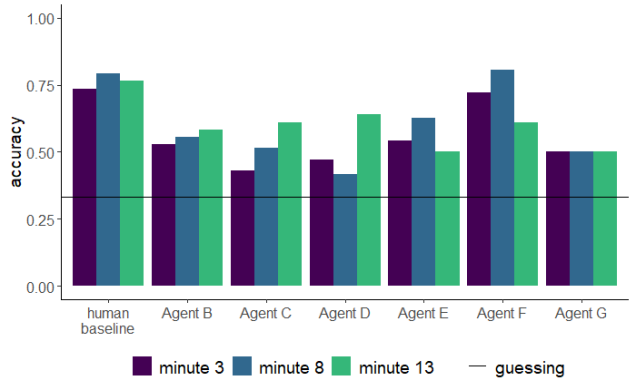


Figure 8: Percent accuracy for inferring marker block semantics (M6), an indicator of false beliefs.

ASI agents were also able to take advantage of information within the trial as it progressed. This suggests that these ASI agents learned something about the structure of the task and team coordination that enabled them to assess performance (M1) and false beliefs (M6) more accurately as the trial progressed. In the case of

performance scores, ASI may have been able to take advantage of decreasing variance in scores as the trial progressed and the diminishing likelihood of accruing more points by rescuing victims. In the case of M6, ASI agents had additional opportunities to observe participant behavior related to marker block placement and movement given others' placement, and therefore allowed ASI agents the opportunity to update prior beliefs on the likely marker block assignment for each participant. Agents did not reliably increase the accuracy of their inferences concerning divergent map information (M3). Analyses by other program researchers indicate that participants often did not use the information provided by maps. Thus, participant planning, navigation, and communication may have held few of the cues that ASI presumably needed to infer that distribution.

## Conclusions

This study developed a rich search and rescue simulation that elicits human taskwork and teamwork. ASI agents successfully used data from this environment to make inferences and predictions that often approached the accuracy of those made by human observers, though ASI and humans used somewhat different data sources (e.g., ASI used message bus traffic and humans used video). The qualitative rationales of the human observers, now under study, may provide insights to refine the design of ASI agents. The rich data provide many opportunities to analyze the relationships between the survey-based variables and action-based variables to further develop reliable and generalizable Machine Theory of Mind (MToM) in the urban search and rescue task environment.

The generalizability of these findings will be tested in planned research. In a 2022 experiment, we will introduce significant perturbations in the task, such as deprivation of communications or changes in task structure or rewards. In research after that, we plan to change the task domain. We predict that ASI will generalize if they develop and maintain an accurate MToT, that is if they are focused not on individual USAR tasks, but on teamwork skills such as leadership, backup behavior, and communication.

Future research will also develop ASI agents that advise teams by leveraging the inferential and predictive abilities enabled by a MToT. That research will evaluate the effects of ASI interventions on team performance, team process, and team member perceptions of the utility and trustworthiness of the ASI designed to aid them.

## Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0130. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

## References

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition* 21(1), 37–46.

Brown, T.A., Perry, S.K.B., Braun, M.T., McCormack, R. Orvis, K.L., & DeCostanza, A.H. (2017, July). A dynamic exploration of multiteam system face-to-face boundary spanning. *Paper presented at the 2017 Annual INGroup Conference, Minneapolis, MN.*

Corral, C., Tatapudi, K., Buchanan, V., Huang, L., and Cooke, N. (in press). Building a Synthetic Task Environment to Support Artificial Social Intelligence Research. *In Proceedings of the International Human Factors and Ergonomics Society Annual Meeting.*

Gopnik, A. and Wellman, H. (1992). Why the child's theory of mind really is a theory. *Mind and Language,* 7:145–171.

Gordon R. (1986). Folk psychology as simulation. *Mind and Language*, 1:158–171.

Huang, L., Freeman, J., Cooke, N., Dubrow, S., Colonna-Romano, J., Wood, M. D., Buchanan, V., & Cauffman, S. J. (2021, June 7). ASIST Study 2 June 2021 Exercises for Artificial Social Intelligence in Minecraft Search and Rescue for Teams. https://doi.org/10.17605/OSF.IO/GXPQ5

Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E., & Matsubara, H. (1997). RoboCup: A Challenge Problem for AI. AI Magazine, 18(1), 73. https://doi.org/10.1609/aimag.v18i1.1276

Lematta, G. J., Coleman, P. B., Bhatti, S. A., Chiou, E. K., McNeese, N. J., Demir, M., & Cooke, N. J. 2019. Developing Human-Robot Team Interdependence in a Synthetic Task Environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63(1), 1503–1507.* https://doi.org/10.1177/1071181319631433

MacMillan, J., Paley, M. J., Levchuk, Y. N., Entin, E. E., Serfaty, D., & Freeman, J. T. 2002. Designing the best team for the task: optimal organizational structures for military missions. M. McNeese, E. Salas, & M. Endsley (Eds.), *New Trends in cooperative activities: System dynamics in complex settings.* San Diego, CA: Human Factors and Ergonomics Society Press.

McNeese, N. J., Cooke, N. J., Fedele, M. A., & Gray, R. 2015. Theoretical and methodical approaches to studying team cognition in sports. *Procedia Manufacturing* 3, 1211-1218.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind*? Behavioral and brain sciences*, 1(4), 515-526.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018, July). Machine theory of mind. *In International conference on machine learning* 4218-4227. PMLR.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13(1), 103–128.

Yurko, R., Matano, F., Richardson, L. F., Granered, N., Pospisil, T., Pelechrinis, K., & Ventura, S. L. 2020. Going deep: models for continuous-time within-play valuation of game outcomes in American football with tracking data. *Journal of Quantitative Analysis in Sports* 16(2), 163-182.